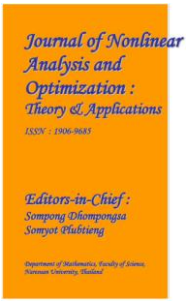
**J**ournal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**




# DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

*[1]K. Vishnu Vardhan Reddy,[2]Ch.Naveen,[3]D.Sudheer,[4]B.Jaipal*
*[1,2,3]IV Year Student,[4]Assistant Professor*
*Department of CSE*
*Visvesvaraya College Of Engineering & Technology,Ibrahimpatnam,Telangana*

**ABSTRACT**

This paper deals with the prediction of Diabetes Disease by performing an analysis of five supervised machine learning algorithms, i.e., K-Nearest Neighbors, Naïve Baye, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

## I. INTRODUCTION

In this day and age, one of the most notorious diseases to have taken the world by storm is Diabetes, which is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin. Due to the many criteria to be taken into consideration for an individual to Harbor this disease, it's detection and prediction might be tedious or sometimes inconclusive. Nevertheless, it isn't impossible to detect it, even at an early stage. Federation- IDF). 79% of the adult population were living in the countries with the low and middle-income groups. It is estimated that by the year 2045 approx. 700 million people will have diabetes (IDF).

Diabetes is increasing day by day in the world because of environmental, genetic factors. The numbers are rising rapidly due to several factors which includes unhealthy foods, physical inactivity and many more. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels, Blood Pressure, etc., play an important role to the contribution of the disease.

In the present we can see that the number of cases is rising every year and there is not slowing down in the active cases. It is a very crucial thing to worry as diabetes has become one of the most dangerous and fastest diseases to take the lives of many individuals around the globe.

Machine Learning is very popular these days as it is used everywhere, where a large amount of data is present, and we need some knowledge from it. Generally, we can categorise the Machine Learning algorithms in two types but not limited to-

- Unsupervised Learning: In unsupervised learning, the information is not labelled and also not trained. Here, we just put the data in action to find some patterns if possible.
- Supervised Learning: In supervised learning, we train the model based on the labels

attached to the information and based on that we classify or test the new data with labels.

With the rise of Machine Learning and its relative algorithms, it has come to light that the significant problems and hindrances in its detection faced earlier, can now be eased with much simplicity, yet, giving a detailed and accurate outcome. As of the modern-day, it is comprehended that Machine Learning has become even more effective and helpful in collaboration with the domain of Medicine. Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual. Such early tries can lead to the inhibition of disease as well as obstruction of permitting the disease to reach a critical degree. The work which will be described in this paper is to perform the diabetes disease prediction using machine learning algorithms for early care of an individual.

## II. LITERATURE SURVEY

### 2.1 EXISTING SYSTEM

In previous, they have used the WEKA tool for data analytics for diabetes disease prediction on Big Data of healthcare. They used the publicly available dataset from UCI and applied different machine learning classifiers on it. The classifiers which they incorporated are Naive Bayes, Support Vector Machine, Random Forest and Simple CART.

Their approach starts with accessing the dataset, preprocess it in Weka tool and then did the 70:30 train and test split for applying different machine algorithms. They did not go with the cross-validation step as it is imperative to get the optimal and accurate results as well.

The authors also used the publicly available dataset named as Pima Indians Diabetes Database for performing their experiment. Their framework of performing the prediction starts with the dataset selection and then with data pre-processing. Once the data was preprocessed, they applied three classification algorithms, i.e., Naïve Bayes, SVM and Decision tree. As they incorporated different evaluation metrics, they did compare the different performance measure and comparatively analyzed the accuracy. The highest accuracy achieved with their experiment was 76.30%. Like they have also not practiced Cross-validation.

### 2.2 DISADVANTAGES OF EXISTING SYSTEM

1). There are no techniques and models for analyzing large scale datasets in the existing system.

2). There is no facility for diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results.

### 2.3 PROPOSED SYSTEM

To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database [4]. This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old. Our proposed model exists of 5 phases which are shown in the proposed system by following Figure.

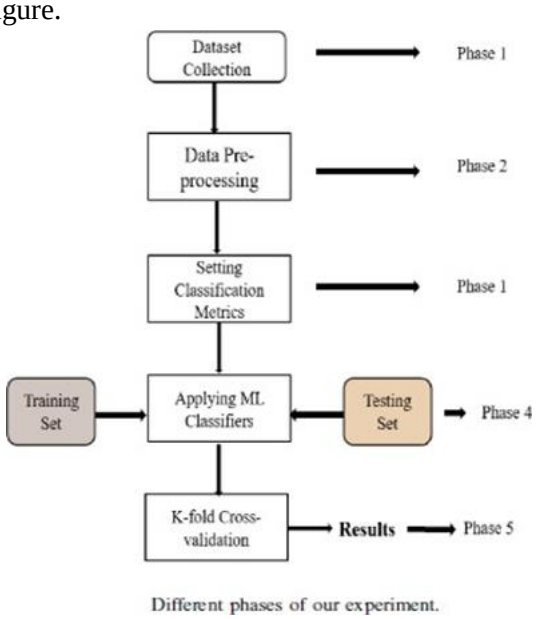



Fig 1. Different phases of our Experiment

### 2.1 ADVANTAGES OF PROPOSED SYSTEM

- The system more effective due to fitting datasets for different ML Models by Applying Machine Learning Algorithms.
- The Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual in the proposed system.

## III. MODULES

**(i) Service Provider-**

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, Find Diabetic Status from Data Set Details, Find Diabetic Ratio on Data Sets, View All Emergency for Diabetic Treatment, Download Trained Data Sets, View Diabetic Ratio Results, View All Remote Users.

**View and Authorize Users**

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

**(ii) Remote User-**

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like POST DIABETIC DATA SETS, SEARCH AND PREDICT DIABETIC STATUS, VIEW YOUR PROFILE.

## IV. ALGORITHMS

The fives supervised machine learning algorithms used in our project are

1) K-NEAREST NEIGHBOURS (KNN)
2) NAIVES BAYES
3) RANDOM FOREST
4) DECISION CLASSIFIERS
5) SUPPORT VECTOR MACHINE

### 4.1 K-NEAREST NEIGHBOURS (KNN)-

Certainly! K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for both classification and regression tasks. It is a type of instance-based learning, also known as lazy learning, as it doesn't create a model during the training phase. Instead, it memorizes the training data and makes predictions based on the similarity of new instances to known instances.

**Here's a breakdown of how the KNN algorithm works:**

**Initialization:**

Store the training dataset. Each data point in the dataset is associated with a class label (in the case of classification) or a numerical value (in the case of regression).

**Input Data:**

When a new, unseen data point is given for prediction, the algorithm identifies its k-nearest neighbors from the training dataset.

**Distance Metric:**

The distance between data points is typically measured using metrics such as Euclidean distance, Manhattan distance, or other distance measures, depending on the nature of the data.

**Voting (for Classification) or Averaging (for Regression):**

For classification, the algorithm counts the occurrences of each class among the k-nearest neighbors and assigns the class label with the majority vote to the new data point.

For regression, it calculates the average of the target values of the k-nearest neighbors and assigns this average as the predicted value for the new data point.

**Choice of 'k':**

The value of 'k' is a crucial parameter that needs to be specified. It represents the number of nearest neighbors to consider. A small 'k' may lead to noisy predictions, while a large 'k' might lead to oversmoothed predictions.

**Key Characteristics:**

- KNN is a non-parametric algorithm, meaning it doesn't make any assumptions about the underlying data distribution.
- It's sensitive to outliers in the data.
- The computational cost of making predictions can be high, especially for large datasets.

### 4.2 NAIVES BAYES-

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an explanation based on the

representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b and RapidMiner 4.6.0). We try above all to understand the obtained results.

**4.3 RANDOM FOREST-**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

**4.4 DECISION CLASSIFIERS-**

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi, we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

**4.5 SUPPORT VECTOR MACHINE-**

In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to genetic algorithms (Gas) or perceptrons, both of which

are widely used for classification in machine learning.. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of Gas and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

## V.SYSTEM DESIGN

### SYSTEM ARCHITECHTURE

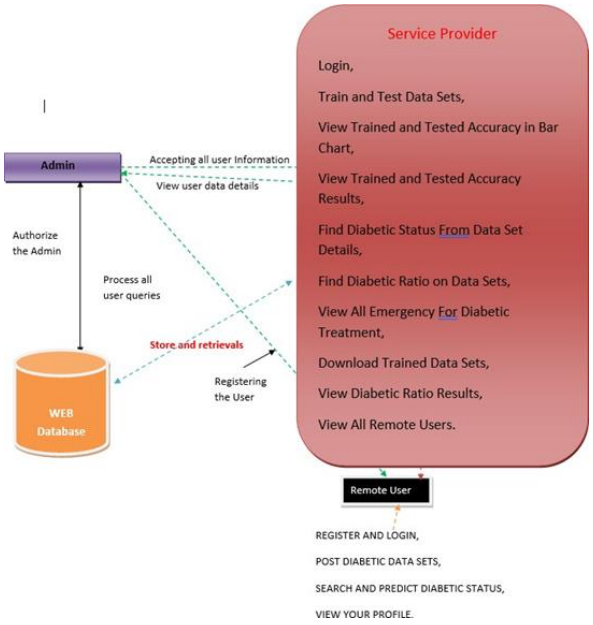



Fig 5.1 Architecture Diagram

## VI. EXECUTION SLIDES

### STARTING THE SERVER IN XAMPP CONTROL PANEL-

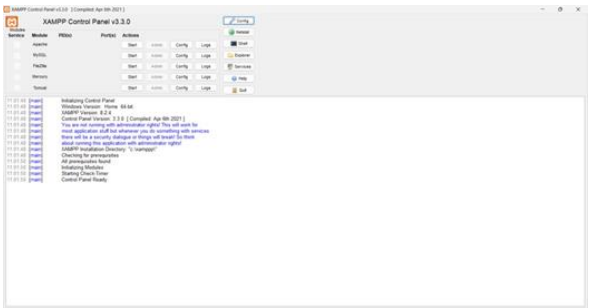



Fig 6.1 Staring the sever in xampp control panel

### USERS, SERVICE PROVIDERS LOGIN AND REGISTRATON-


Fig 6.2 Users,Service Providers Login and Registration

### USER REGISTRATION PAGE-


Fig 6.3 User Registration Page

### USER DATASET UPLOADING PAGE-


Fig 6.4 User Dataset Uploading page

### SERVICE PROVIDER LOGIN PAGE-


Fig 6.5 Service Provider Login Page

### DIABETES STATUS FROM DATASET DETAILS-

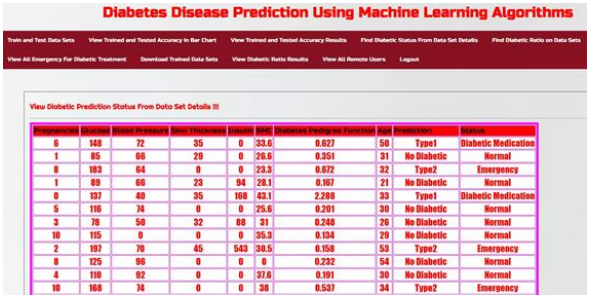



Fig 6.6 Diabetes status from Dataset details

**VIEW TRAINED AND TESTED DATA IN PIE CHART-**

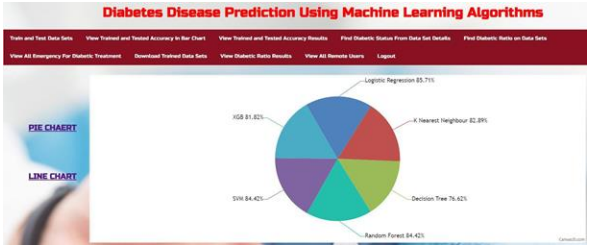



Fig 6.7 View trained and tested data in Pie chart

**VIEW TRAINED AND TESTED DATA IN BAR CHART-**

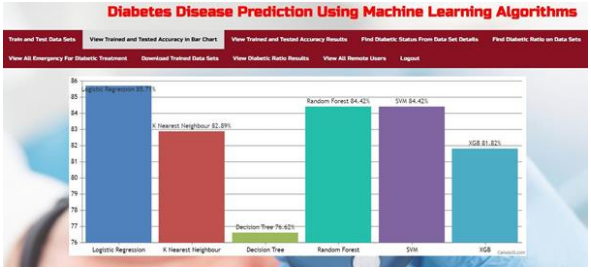



Fig 6.8 View trained and tested data in bar chart

## VII. CONCLUSION

One of the significant impediments with the progression of technology and medicine is the early detection of a disease, which is in this case, diabetes. However, in this study, systematic efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 76% using the K-Nearest Neighbours classifiers. With this being said, it is hopeful that we can implement this model into a system to predict other deadly diseases as well. There can be room for further improvement for the automation of the analysis of diabetes or any other disease in the future.

In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. We will be incorporating more Machine Learning and Deep learning models for achieving better results as well

## FUTURE SCOPE


Fig : Future Scope

Expanding the scope of Diabetes Prediction by incorporating Deep Learning techniques for more Accurate and Timely Predictions, and developing a User-Friendly Application for widespread accessibility and usability

In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results.

## REFERENCES

1. P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga,N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova,J. E. Shaw, D. Bright, and R.Williams, “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition,” Diabetes Research and Clinical Practice, vol. 157, p. 107843, 2019.
2. A. Mir and S. N. Dhage, “Diabetes disease prediction using machine learning on big data of healthcare,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–6.
3. D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” Procedia Computer Science, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918308548

4. J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, “Using the adap learning algorithm to forcast the onset of diabetes mellitus,” Proceedings - Annual Symposium on Computer Applications in Medical Care, vol. 10, 11 1988.
5. P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.
6. Wes McKinney, “Data Structures for Statistical Computing in Python,” in Proceedings of the 9th Python in Science Conference, St´efan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.
7. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers,P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk,M. Brett, A. Haldane, J. F. del R’ıo, M. Wiebe, P. Peterson,P. G’erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser,H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-649-2
8. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg,J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and´ Edouard Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, p. 28252830, 2011.